Dimension Reduction in Regression Estimation with Nearest Neighbor

Benoît CADRE*, Qian DONG

IRMAR, ENS Cachan Bretagne, CNRS, UEB
Campus de Ker Lann
Avenue Robert Schuman, 35170 Bruz, France
cadre@bretagne.ens-cachan.fr

Abstract

In regression with a high-dimensional predictor vector, dimension reduction methods aim at replacing the predictor by a lower dimensional version without loss of information on the regression. In this context, the so-called central mean subspace is the key of dimension reduction. The last two decades have seen the emergence of many methods to estimate the central mean subspace. In this paper, we go one step further, and we study the performances of a k-nearest neighbor type estimate of the regression function, based on an estimator of the central mean subspace. The estimate is first proved to be consistent. Improvement due to the dimension reduction step is then observed in term of its rate of convergence. All the results are distributions-free. As an application, we give an explicit rate of convergence using the SIR method.

Index Terms — Dimension Reduction; Central Mean Subspace; Nearest Neighbor Method; Semiparametric Regression; SIR Method.

AMS 2000 Classification — 62H12; 62G08.

1 Introduction

In a full generality, the goal of regression is to infer about the conditional law of the response variable Y given the $\mathbb{R}^p$-valued predictor X. Many different methods

*corresponding author

have been developped to adress this issue. In the present paper, we consider sufficient dimension reduction which is a body of theory and methods for reducing the dimension of X while preserving information on the regression (see Li, 1991, 1992 and Cook and Weisberg, 1991). Basically, the idea is to replace the predictor with its projection onto a subspace of the predictor space, without loss of information on the conditional distribution of Y given X. Several methods have been introduced to estimate this subspace: sliced inverse regression (SIR; Li, 1991), sliced average variance estimation (SAVE; Cook and Weisberg, 1991), average derivative estimation (ADE; Härdle and Stoker, 1989), ... See also the paper by Cook and Weisberg (1999) who gives an introductory account of studying regression via these methods.

Even if the methods above give a complete picture of the dependence of Yon X, certain characteristics of the conditional distribution may often be of special interest. In particular, regression is often understood to imply a study of the conditional expectation $\mathbb{E}[Y|X]$. Subsequently, the response variable Y is a univariate and integrable random variable. Following the ideas developped for the conditional distribution, Cook and Li (2002) introduced the *central mean subspace* that will be of great interest for the paper. Let us recall the definition. For a matrix $\Lambda \in \mathscr{M}_p(\mathbb{R})$, denote by $\mathbf{S}(\Lambda)$ the space spanned by the columns of Λ. Here, $\mathscr{M}_p(\mathbb{R})$ stands for the set of $p \times p$-matrices with real coefficients. Letting Λ^T the transpose matrix of Λ, we say that $\mathbf{S}(\Lambda)$ is a *mean dimension-reduction subspace* if

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|\Lambda^T X], \tag{1.1}$$

that is, if the projection of the predictor onto $\mathbf{S}(\Lambda)$ has no influence on the regression. When the intersection of all dimension-reduction subspaces itself is a dimension-reduction subspace, it is defined as the central mean subspace and is denoted by $\mathscr{S}_{\mathbb{E}[Y|X]}$. With this respect, a matrix Λ that spans the central mean subspace is called a *candidate matrix*. Hence the central mean subspace, which exists under mild conditions (see Cook, 1994, 1996, 1998), is the target of sufficient dimension reduction for the mean response $\mathbb{E}[Y|X]$. Various methods have been developped to estimate $\mathscr{S}_{\mathbb{E}[Y|X]}$, among with principle Hessian direction (pHd; Li, 1992), iterative Hessian transformation (IHT; Cook and Li, 2002), minimum average variance estimation (MAVE; Xia et al, 2002). Discussions, improvements and relevant papers can be found in Zhu and Zeng (2006), Ye and Weiss (2003) or Cook and Ni (2005).

Regarding the regression estimation problem in a nonparametric setting, the aim of the dimension-reduction methods is to overcome the *curse of dimensionality* -which roughly says that the rate of convergence of any estimator decreases as p grows- by accelerating the rate of convergence. Indeed, assuming (1.1), it is naturally expected that the rate of convergence of any estimator will depend on $\mathrm{rank}(\Lambda)$ instead of p, since $\Lambda^T X$ lies in a vector space of dimension $\mathrm{rank}(\Lambda)$. In general, $\mathrm{rank}(\Lambda)$ is much smaller than p, hence the rate of convergence in the estimation of $\mathbb{E}[Y|X]$ may be considerably improved. For this estimation problem, we shall use the so-called k-nearest neighbor method (NN), which is one of the most studied method in nonparametric regression estimation since it provides efficient and tractable estimators (e.g., see the monography by Györfi et al, 2002, and the references therein). As far as we know, similar studies in a dimension-reduction setting were only been carried out for particular models, such as additive models or projection pursuits for instance. We refer the reader to Chapter 22 in the book by Györfi et al (2002) for a complete list of references on the subject.

In the present paper, we adress the problem of estimating the conditional expectation $\mathbb{E}[Y|X]$ based on a sequence $(X_1,Y_1),\cdots,(X_N,Y_N)$ of i.i.d. copies of (X,Y). Assuming the existence of a mean dimension-reduction subspace as in (1.1), we first construct in Section 2 the k-NN type estimator based on an estimate $\hat{\Lambda}$ of Λ. Roughly speaking, it is defined as the k-NN regression estimate drawn from the $(\hat{\Lambda}X_i,Y_i)$'s. In a distribution-free setting, we prove consistency of the estimator (Theorem 2.1) and we show that the rate of convergence essentially depends on $\mathrm{rank}(\Lambda)$ (Theorem 2.2). In particular, up to the terms induced by the dimension-reduction methodology, we recover the usual optimal rate when the predictor belongs to $\mathbb{R}^{\mathrm{rank}(\Lambda)}$. Section 3 is devoted to the term induced by the dimension-reduction method: in a general setting, we propose and study the performances (convergence and rate) of a numerically robust estimator. As an example, we consider in Section 4 the case where the candidate matrix is constructed via the SIR method. All the proofs are postponed to the last three sections.

2 Fast regression estimation

2.1 The estimator

Throughout this section, we shall assume the following assumption.

BASIC ASSUMPTION: *there exists* $\Lambda \in \mathscr{M}_p(\mathbb{R})$ *such that* $\mathbf{S}(\Lambda^T)$ *is a mean dimension-reduction subspace, i.e.*

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|\Lambda X].$$

Note that we have written "Λ" instead of the usual "Λ^T" in the conditional expectation. This choice is for notational simplicity since, in this section, we only have to deal with Λ.

The estimation of the regression function requires to first estimate the matrix Λ and then to estimate the regression function r defined by

$$r(x) = \mathbb{E}[Y|\Lambda X = x], \quad x \in \mathbb{R}^p.$$

To reach this goal, we assume throughout the paper that the sample size N is even, with $N = 2n$. We split the dataset into two sub-samples: the n first data $(X_1, Y_1), \cdots, (X_n, Y_n)$ are used to estimate the matrix Λ, whereas the last ones $(X_{n+1}, Y_{n+1}), \cdots, (X_{2n}, Y_{2n})$ are used to estimate the body of the regression function r.

For the first estimation problem, we assume in this section that we have at hand an estimate $\hat{\Lambda}$ of Λ, constructed with the observations $(X_1, Y_1), \cdots, (X_n, Y_n)$. We refer to Sections 3 and 4 for an efficient and tractable way to estimate Λ. We now explain the nearest neighbour method that will be introduced to estimate the function r (for more information on the NN-method, we refer the reader to Chapter 6 of the monography by Györfi et al, 2002). For all $i = n+1, \cdots, 2n$, we let

$$\hat{X}_i = \hat{\Lambda} X_i.$$

Then, if $x \in \mathbb{R}^p$, we reorder the data $(\hat{X}_{n+1}, Y_{n+1}), \cdots, (\hat{X}_{2n}, Y_{2n})$ according to increasing values of $\{\|\hat{X}_i - x\|, i = n+1, \cdots, 2n\}$, where $\|.\|$ stands for the Schur norm of any vector or matrix. The reordered data sequence is denoted by:

$$(\hat{X}_{(1)}(x), Y_{(1)}(x)), (\hat{X}_{(2)}(x), Y_{(2)}(x)), \cdots, (\hat{X}_{(n)}(x), Y_{(n)}(x)),$$

which means that

$$\|\hat{X}_{(1)}(x) - x\| \leq \|\hat{X}_{(2)}(x) - x\| \leq \cdots \leq \|\hat{X}_{(n)}(x) - x\|.$$

In this approach, $\hat{X}_{(i)}(x)$ is called the i-th NN of x. Note that if $\hat{X}_i$ and $\hat{X}_j$ are equidistant from x, i.e. $\|\hat{X}_i - x\| = \|\hat{X}_j - x\|$, then we have a tie. As usual, we then

declare $\hat{X}_i$ closer to x than $\hat{X}_j$ if $i < j$. We now let $k = k(n) \leq n$ be an integer and for all $i = n+1, \cdots, 2n$, we set:

$$W_i(x) = \begin{cases} 1/k & \text{if } \hat{X}_i \text{ is among the } k\text{-NN of } x \text{ in } \{\hat{X}_{n+1}, \cdots, \hat{X}_{2n}\}; \\ 0 & \text{elsewhere.} \end{cases}$$

Observe that we have $\sum_{i=n+1}^{2n} W_i(x) = 1$. With this respect, the estimate $\hat{r}$ of r is then defined by:

$$\hat{r}(x) = \sum_{i=n+1}^{2n} W_i(x) Y_i = \frac{1}{k} \sum_{i=1}^{k} Y_{(i)}(x), \ x \in \mathbb{R}^p.$$

From a computational point of view, the complexity of the calculation algorithm of $\hat{r}(x)$ is $O(n \ln n)$ in mean, using a random Quick-Sort Algorithm.

2.2 Behavior of $\hat{r}$

In the sequel, (X,Y) is independent of the whole sample and with the same distribution as (X_1,Y_1). Observe that our results are distribution-free; in particular, we do not assume that the law of (X,Y) has a density. The first result, whose proof is deferred to Section 5, establishes a consistency property for the estimator $\hat{r}(\hat{\Lambda}X)$.

Theorem 2.1. *Assume that Y is bounded. If $k \to \infty$, $k/n \to 0$ and $\hat{\Lambda} \xrightarrow{\mathbb{P}} \Lambda$, then:*

$$\hat{r}(\hat{\Lambda}X) \xrightarrow{\mathbb{L}^2} \mathbb{E}[Y|X].$$

Therefore, we assume in the following that $k/n \to 0$. Recall that the consistency assumption $\hat{\Lambda} \xrightarrow{\mathbb{P}} \Lambda$ holds for the standard dimension reductions methodologies, as we shall see in Sections 3 and 4.

We now turn to the study of the rate of convergence. Recall that the function r is lipschitz if there exists $L > 0$ such that for all $x_1, x_2 \in \mathbb{R}^p$:

$$|r(x_1) - r(x_2)| \leq L\|x_1 - x_2\|.$$

Because we deal with the estimation of $\mathbb{E}[Y|\Lambda X]$, it is naturally expected that the convergence rate in Theorem 2.1 depends on the dimension of the vector space spanned by the matrix Λ. In the sequel, d stands for the rank of Λ, and we also denote by $\hat{d}$ an estimator such that $\hat{d} = \text{rank}(\hat{\Lambda})$. Section 6 is devoted to the proof of the following result:

Theorem 2.2. *Assume that X and Y are bounded. If r is lipschitz and $d \geq 3$, there exists a constant $C > 0$ such that*

$$\mathbb{E}\left(\hat{r}(\hat{\Lambda}X) - \mathbb{E}[Y|X]\right)^2 \leq \frac{C}{k} + C\left(\frac{k}{n}\right)^{2/d} + C\,\mathbb{E}\|\hat{\Lambda} - \Lambda\|^2 + \mathbb{P}(\hat{d} > d).$$

Remark 2.3. *When $d \leq 2$, under the additional conditions of Problem 6.7 in the book by Györfi et al (2002), a slight adaptation of the proof of Theorem 2.2 enables us to derive the same convergence rate.*

Observe that the global error is decomposed into two terms: first, the classical error term

$$\frac{C}{k} + C\left(\frac{k}{n}\right)^{2/d}$$

in nonparametric regression estimation using k-NN, but when the predictor belongs to $\mathbb{R}^d$ (see Chapter 6 in the book by Györfi et al, 2002) ; seconds, the term

$$C\,\mathbb{E}\|\hat{\Lambda} - \Lambda\|^2 + \mathbb{P}(\hat{d} > d)$$

induced by the dimension-reduction method. We shall concentrate on this term in the next two sections.

Note also that in this result, the best choice of k, namely $k = n^{2/(2+d)}$, gives the following bound:

$$\mathbb{E}\left(\hat{r}(\hat{\Lambda}X) - \mathbb{E}[Y|X]\right)^2 \leq 2C\,n^{-2/(d+2)} + C\,\mathbb{E}\|\hat{\Lambda} - \Lambda\|^2 + \mathbb{P}(\hat{d} > d).$$

Hence, up to the last two terms, our nearest neighbor estimate achieves the usual optimal rate in regression estimation, but when the predictor belongs to $\mathbb{R}^d$ (see Ibragimov and Khasminskii, 1981 or Györfi et al, 2002). With this result, one may quantify the positive effects of the dimension reduction step, that are measured in term of the rate of convergence.

Next section is dedicated to the construction and estimation of Λ in a general setting.

3 General dimension reduction methodology

3.1 Construction of Λ

Papers dealing about dimension reduction primarily focus on the determination of a candidate matrix $M \in \mathscr{M}_p(\mathbb{R})$ such that the central mean subspace is spanned by the columns of M, i.e. $\mathbf{S}(M) = \mathscr{S}_{\mathbb{E}[Y|X]}$. Observe that the matrix M is symmetric for the standard dimension-reduction methodologies. We shall see in the next section an explicit description of M with the SIR method. Note that the matrix M is in some sense minimal because it spans the smallest mean dimension-reduction subspace.

In this section, the matrix Λ of Section 2 will be constructed from a candidate matrix with a spectral decomposition. There are two main reasons for this: first, it automatically gives the effective directions of the reduced space; seconds, the thresholding procedure of the empirical eigenvalues developped below is robust from a numerical point of view.

Here, we only have to assume that $M \in \mathscr{M}_p(\mathbb{R})$ is a symmetric matrix such that $\mathbf{S}(M)$ is a mean dimension-reduction subspace, i.e.

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|M^T X].$$

We let $\text{rank}(M) = d$. Furthemore, we denote by $\lambda_1, \cdots, \lambda_p$ the eigenvalues of M indexed as follows:

$$\lambda_1 \geq \cdots \geq \lambda_p.$$

Set now $v_1, \cdots, v_p$ the normalized eigenvectors associated with $\lambda_1, \cdots, \lambda_p$, and $\ell_1 < \cdots < \ell_d$ the integers such that $\lambda_{\ell_j} \neq 0$ for all $j = 1, \cdots, p$. Recall that $v_1, \cdots, v_p$ are orthogonal vectors. In the particular case where M is positive definite, $\ell_i = i$. Let $\mathbf{O}$ be the null-vector in $\mathbb{R}^p$. The matrix Λ of Section 2 is defined by:

$$\Lambda^T = \begin{pmatrix} v_{\ell_1} & \cdots & v_{\ell_d} & \mathbf{O} & \cdots & \mathbf{O} \end{pmatrix},$$

so that $\text{rank}(\Lambda) = d$ and $\mathbb{E}[Y|X] = \mathbb{E}[Y|\Lambda X]$ because $\mathbf{S}(M) = \mathbf{S}(\Lambda^T)$. In particular, the basic assumption of Section 2 holds.

We also assume that we have at hand the estimator $\hat{M} \in \mathscr{M}_p(\mathbb{R})$ of M, constructed with the n first data $(X_1, Y_1), \cdots, (X_n, Y_n)$. We suppose that $\hat{M}$ is a symmetric

matrix with real coefficients, and we denote by $\hat{\lambda}_1, \cdots, \hat{\lambda}_p$ the eigenvalues indexed as follows:

$$\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p,$$

and by $\hat{v}_1, \cdots, \hat{v}_p$ the corresponding normalized eigenvectors. A natural -and numerically robust- estimator $\hat{d}$ of d is then obtained by thresholding the eigenvalues:

$$\hat{d} = \sum_{j=1}^{p} \mathbf{1}\{|\hat{\lambda}_j| \geq \tau\},$$

where the threshold τ is some positive real number with $\tau \leq 1$, to be specified latter. Let $\hat{\ell}_1 < \cdots < \hat{\ell}_{\hat{d}}$ be the integers such that $|\hat{\lambda}_{\hat{\ell}_j}| \geq r$ for all $j = 1, \cdots, \hat{d}$. Then, we put:

$$\hat{\Lambda}^T = \begin{pmatrix} \hat{v}_{\hat{\ell}_1} & \cdots & \hat{v}_{\hat{\ell}_{\hat{d}}} & \mathbf{O} & \cdots & \mathbf{O} \end{pmatrix},$$

and we observe that $\text{rank}(\hat{\Lambda}) = \hat{d}$.

3.2 Rate of convergence

It is an easy task to prove that if $\hat{M} \xrightarrow{\mathbb{P}} M$, then $\hat{\Lambda} \xrightarrow{\mathbb{P}} \Lambda$. Hence by Theorem 2.1, if Y is bounded, we have:

$$\hat{r}(\hat{\Lambda} X) \xrightarrow{\mathbb{L}^2} \mathbb{E}[Y|X],$$

provided $k \to \infty$ and $k/n \to 0$. This subsection is dedicated to the rate of convergence in the above convergence result.

As seen in Theorem 2.2, we need to give bounds for both terms $\mathbb{P}(\hat{d} > d)$ and $\mathbb{E}\|\hat{\Lambda} - \Lambda\|^2$. The bounds are given in Lemmas 7.1 and 7.2 in Section 7. As an application of Corollary 2.2, we immediately deduce the following result:

Corollary 3.1. *Assume that X and Y are bounded, $d \geq 3$ and r is lipschitz. If the non-null eigenvalues of M have multiplicity 1, then there exists a constant $C > 0$ such that*

$$\mathbb{E}\left(\hat{r}(\hat{\Lambda} X) - \mathbb{E}[Y|X]\right)^2 \leq \frac{C}{k} + C\left(\frac{k}{n}\right)^{2/d} + \frac{C}{\tau^2}\mathbb{E}\|\hat{M} - M\|^2.$$

Next section is dedicated to the case where M is constructed with the SIR method. In this context, we can give a bound for $\mathbb{E}\|\hat{M} - M\|^2$, hence an explicit rate of convergence of $\hat{r}(\hat{\Lambda} X)$ to $\mathbb{E}[Y|X]$.

4 Application with the SIR method

The goal of this section is to apply Corollary 3.1 when the candidate matrix M is constructed with some dimension-reduction method. It appears that for each dimension-reduction method (SIR, ADE, MAVE, ...), the estimator $\hat{M}$ of M is such that $\sqrt{n}(\hat{M}-M)$ converges in distribution. However, in view of an application of Corollary 3.1, we need a bound for the quantity $\mathbb{E}\|\hat{M}-M\|^2$. Each dimension-reduction method need a specific process, and an exhaustive study of all processes is beyond the scope of the paper.

Hence, we have chosen to study the case where M is constructed with SIR, since it is one of the most popular and powerfull dimension-reduction method, and because it is the subject of many recent papers (see for instance the papers by Saracco, 2005, and Zhu et al, 2006, and the references therein).

In this section, we assume that X and Y are bounded. For simplicity, we also assume that X is standard, i.e. X has mean 0 and variance matrix Id. With the SIR method, the candidate matrix M of Section 3, further denoted M_{SIR}, is the symmetric matrix defined by:

$$M_{\mathrm{SIR}} = \mathrm{cov}\left(\mathbb{E}[X|Y]\right) = \mathbb{E}\left(\mathbb{E}[X|Y]\mathbb{E}[X|Y]^T\right).$$

In view of an application of Corollary 3.1, we assume throughout that

$$\mathbf{S}(M_{\mathrm{SIR}}) = \mathscr{S}_{Y|X},$$

where $\mathscr{S}_{Y|X}$ stands for the *central subspace* of Y given X (e.g. see Li, 1991). We refer to the papers by Li (1991) and Hall and Li (1993) for discussions on this assumption, as well as sufficient conditions on the model that ensures this property. In particular,

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|M_{\mathrm{SIR}}^T X],$$

hence we are in position to apply the results of Section 3.

Let us introduce the partition $\{I(h), h = 1, \cdots, H\}$ of the support of Y, such that each *slice* $I(h)$ (shorten as h) is an interval with length κ/H for some $\kappa > 0$, and moreover:

$$p_h = \mathbb{P}(Y \in I(h)) > 0.$$

With this respect, a natural estimator for the SIR matrix M_{SIR} is

$$\hat{M}_{\text{SIR}} = \sum_{h=1}^{H} \hat{p}_h \hat{m}_h \hat{m}_h^T,$$

where for any slice h:

$$\hat{p}_h = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i \in I(h)\} \quad \text{and} \quad \hat{m}_h = \frac{1}{n\hat{p}_h}\sum_{i=1}^{n} X_i \mathbf{1}\{Y_i \in I(h)\}.$$

We now denote by m_h the theoretical counterpart of $\hat{m}_h$, i.e. $m_h = \mathbb{E}[X|Y \in I(h)]$, and by M'_{SIR} the matrix:

$$M'_{\text{SIR}} = \sum_{h=1}^{H} p_h m_h m_h^T.$$

It is an easy exercise to prove that

$$\mathbb{E}\|\hat{M}_{\text{SIR}} - M'_{\text{SIR}}\|^2 \leq C\frac{H}{n}, \tag{4.1}$$

for some constant $C > 0$ that does not depend on n and H. Hence in the estimation of M_{SIR} by $\hat{M}_{\text{SIR}}$, the bound on the variance term does not need additional assumptions. The biais term $\|M'_{\text{SIR}} - M_{\text{SIR}}\|$, however, has to be handled with care. In the sequel, r_{inv} stands for the *inverse regression function*, that is:

$$r_{\text{inv}}(y) = \mathbb{E}[X|Y = y].$$

We observe that for each slice h:

$$m_h = \frac{1}{p_h}\mathbb{E}X\mathbf{1}\{Y \in I(h)\} = \frac{1}{p_h}\mathbb{E}\, r_{\text{inv}}(Y)\mathbf{1}\{Y \in I(h)\}.$$

Hence, provided r_{inv} is Lipschitz, one obtains:

$$\|M'_{\text{SIR}} - \sum_{h=1}^{H} p_h r_{\text{inv}}(c_h) r_{\text{inv}}(c_h)^T\| \leq \frac{C}{H}, \tag{4.2}$$

for some constant $C > 0$, and where the c_h's are contained in the $I(h)$'s. Moreover, we observe that M_{SIR} can be written as

$$M_{\text{SIR}} = \sum_{h=1}^{H} \mathbb{E}\,\mathbf{1}\{Y \in I(h)\} r_{\text{inv}}(Y) r_{\text{inv}}(Y)^T.$$

Therefore,

$$\|M_{\mathrm{SIR}} - \sum_{h=1}^{H} p_h r_{\mathrm{inv}}(c_h) r_{\mathrm{inv}}(c_h)^T\| \leq \frac{C}{H}, \tag{4.3}$$

for some constant $C > 0$. Under the Lipschitz assumption on r_{inv}, we thus get from (4.1), (4.2) and (4.3):

$$\|\hat{M}_{\mathrm{SIR}} - M_{\mathrm{SIR}}\|^2 \leq C\left(\frac{H}{n} + \frac{1}{H^2}\right),$$

for some constant $C > 0$.

In the sequel, Λ_{SIR} (resp. $\hat{\Lambda}_{\mathrm{SIR}}$) is constructed with the matrix $M = M_{\mathrm{SIR}}$ (resp. $\hat{M} = \hat{M}_{\mathrm{SIR}}$) as in Section 3.1 and d is the rank of M_{SIR}. For the construction of the estimate, one has to choose the values of the parameters H (the number of slices), τ (the thresholding parameter of the eigenvalues) and k (the number of NN). With the above choices:

$$H = n^{1/3}, \quad \tau = n^{1/6} \quad \text{and} \quad k = n^{2/(2+d)},$$

we immediatly deduce from Corollary 3.1 our last result.

Corollary 4.1. *Assume that $d \geq 3$. If r and r_{inv} are Lipschitz, and if the non-null eigenvalues of M_{SIR} have multiplicity 1, then there exists a constant $C > 0$ such that*

$$\mathbb{E}\left(\hat{r}(\hat{\Lambda}_{\mathrm{SIR}} X) - \mathbb{E}[Y|X]\right)^2 \leq C n^{-2/(2+d)}.$$

Hence, we recover the usual optimal rate when the predictor vector belongs to a d-dimensional vector space.

5 Proof of Theorem 2.1

5.1 Preliminaries

For simplicity, we assume that $|Y| \leq 1$. We let $\hat{X} = \hat{\Lambda} X$, $\tilde{X} = \Lambda X$ and, for all $i = n+1, \cdots, 2n$:

$$\tilde{X}_i = \Lambda X_i.$$

Lemma 5.1. *If $k/n \to 0$ and $\hat{\Lambda} \xrightarrow{\mathbb{P}} \Lambda$, then*

$$\hat{X}_{(k)}(\hat{X}) - \hat{X} \xrightarrow{\mathbb{P}} 0.$$

Proof In the proof, μ stands for the distribution of X. Let $\varepsilon > 0$. Since X is independent from the sample and distributed according to μ, we have the following equality:

$$\mathbb{P}(\|\hat{X}_{(k)}(\hat{X}) - \hat{X}\| > \varepsilon) = \int_{\mathbb{R}^p} \mathbb{P}(\|\hat{X}_{(k)}(\hat{\Lambda}x) - \hat{\Lambda}x\| > \varepsilon)\, \mu(\mathrm{d}x).$$

Then, according to the Lebesgue domination Theorem, one only needs to prove that for all x in the support of μ:

$$\mathbb{P}(\|\hat{X}_{(k)}(\hat{\Lambda}x) - \hat{\Lambda}x\| > \varepsilon) \longrightarrow 0.$$

Observe now that for all x:

$$\mathbb{P}(\|\hat{X}_{(k)}(\hat{\Lambda}x) - \hat{\Lambda}x\| > \varepsilon) = \mathbb{P}\left(\sum_{i=n+1}^{2n} \mathbf{1}\{\|\hat{X}_i - \hat{\Lambda}x\| \leq \varepsilon\} < k \right).$$

Let $a > \|\Lambda\|$. If $\|\hat{\Lambda} - \Lambda\| \leq a$ and $\|\hat{X}_i - \hat{\Lambda}x\| \leq \varepsilon$, we then have:

$$\begin{aligned} \|\tilde{X}_i - \Lambda x\| &\leq \|(\Lambda - \hat{\Lambda})X_i\| + \|\hat{X}_i - \hat{\Lambda}x\| + \|(\hat{\Lambda} - \Lambda)x\| \\ &\leq a(\|X_i\| + \|x\|) + \varepsilon. \end{aligned}$$

Therefore,

$$\begin{aligned} &\mathbb{P}(\|\hat{X}_{(k)}(\hat{\Lambda}x) - \hat{\Lambda}x\| > \varepsilon) \qquad (5.1) \\ &\leq \mathbb{P}\left(\sum_{i=n+1}^{2n} \mathbf{1}\{\|\tilde{X}_i - \Lambda x\| \leq a(\|X_i\| + \|x\|) + \varepsilon\} < k \right) + \mathbb{P}(\|\hat{\Lambda} - \Lambda\| > a). \end{aligned}$$

According to the strong law of large numbers:

$$\frac{1}{n} \sum_{i=n+1}^{2n} \mathbf{1}\{\|\tilde{X}_i - \Lambda x\| \leq a(\|X_i\| + \|x\|) + \varepsilon\} \xrightarrow{\text{a.s}} \mathbb{P}\left(\|\tilde{X} - \Lambda x\| \leq a(\|X\| + \|x\|) + \varepsilon\right).$$

Assume that the latter quantity equals 0. Then, we have a.s.

$$\|\tilde{X} - \Lambda x\| > a(\|X\| + \|x\|) + \varepsilon.$$

But this is impossible since $\|\tilde{X} - \Lambda x\| \leq \|\Lambda\|(\|X\| + \|x\|)$ and $a > \|\Lambda\|$. As a consequence,

$$\mathbb{P}\left(\|\tilde{X} - \Lambda x\| \leq a(\|X\| + \|x\|) + \varepsilon\right) \neq 0,$$

and, since $k/n \to 0$, we obtain

$$\mathbb{P}\left(\sum_{i=n+1}^{2n} \mathbf{1}\{\|\tilde{X}_i - \Lambda x\| \leq a(\|X_i\| + \|x\|) + \varepsilon\} < k \right) \longrightarrow 0.$$

By assumption, $\mathbb{P}(\|\hat{\Lambda} - \Lambda\| > a) \to 0$ so that by (5.1),

$$\mathbb{P}(\|\hat{X}_{(k)}(\hat{\Lambda} x) - \hat{\Lambda} x\| > \varepsilon) \longrightarrow 0,$$

hence the lemma. □

Lemma 5.2. *Let $\varphi : \mathbb{R}^p \to \mathbb{R}$ be a uniformly continuous function such that $0 \leq \varphi \leq 1$. If $k/n \to 0$ and $\hat{\Lambda} \xrightarrow{\mathbb{P}} \Lambda$, we have:*

$$\mathbb{E} \sum_{i=n+1}^{2n} W_i(\hat{X}) \left(\varphi(\tilde{X}) - \varphi(\tilde{X}_i)\right)^2 \longrightarrow 0.$$

Proof For all $K > 0$, we let $X_K = X\mathbf{1}\{\|X\| \leq K\}$. Then, we note $\hat{X}_K = \hat{\Lambda} X_K$, $\tilde{X}_K = \Lambda X_K$ and similarly for $\hat{X}_{i,K}$ and $\tilde{X}_{i,K}$. Moreover, $W_{i,K}$ is defined as W_i, but with the $\hat{X}_{i,K}$'s instead of the $\hat{X}_i$'s (see Section 2.1). A moment's thought reveals that, since $\sum_{i=n+1}^{2n} W_{i,K}(\hat{X}_K) = 1$:

$$\begin{aligned}
&\mathbb{E} \sum_{i=n+1}^{2n} W_i(\hat{X}) \left(\varphi(\tilde{X}) - \varphi(\tilde{X}_i)\right)^2 \\
&= \mathbb{P}(\|X\| < K)^n \, \mathbb{E} \sum_{i=n+1}^{2n} W_{i,K}(\hat{X}_K) \left(\varphi(\tilde{X}_K) - \varphi(\tilde{X}_{i,K})\right)^2 + R_K,
\end{aligned}$$

where R_K is a positive real number that satisfies $\sup_n R_K \to 0$ as $K \to \infty$. Therefore, one only needs to prove that for all $K > 0$, one has :

$$\mathbb{E} \sum_{i=n+1}^{2n} W_{i,K}(\hat{X}_K) \left(\varphi(\tilde{X}_K) - \varphi(\tilde{X}_{i,K})\right)^2 \longrightarrow 0. \tag{5.2}$$

We now proceed to prove this property.

Fix $K > 0$ and $\varepsilon > 0$. There exists $r > 0$ such that $|\varphi(x_1) - \varphi(x_2)| \leq \varepsilon$ provided $x_1, x_2 \in \mathbb{R}^p$ satisfy $\|x_1 - x_2\| \leq r$. Since φ is bounded by 1 and $\sum_{i=n+1}^{2n} W_{i,K}(\hat{X}_K) =$

1, we have:

$$\mathbb{E}\sum_{i=n+1}^{2n} W_{i,K}(\hat{X}_K)\left(\varphi(\tilde{X}_K)-\varphi(\tilde{X}_{i,K})\right)^2$$
$$\leq \varepsilon^2 + \mathbb{E}\sum_{i=n+1}^{2n} W_{i,K}(\hat{X}_K)\mathbf{1}\{\|\tilde{X}_K-\tilde{X}_{i,K}\|>r\}. \tag{5.3}$$

Hence, one only needs to prove that the rightmost term tends to 0. If $\|\hat{\Lambda}-\Lambda\| \leq r/(4K)$ and $\|\tilde{X}_{i,K}-\tilde{X}_K\|>r$, then:

$$\|\hat{X}_K-\hat{X}_{i,K}\| \geq \|\tilde{X}_K-\tilde{X}_{i,K}\| - \|\hat{\Lambda}-\Lambda\|(\|X_K\|+\|X_{i,K}\|) \geq \frac{r}{2},$$

because $\|X_K\| \leq K$ and $\|X_{i,K}\| \leq K$. Consequently,

$$\begin{aligned}
&\mathbb{E}\sum_{i=n+1}^{2n} W_{i,K}(\hat{X}_K)\mathbf{1}\{\|\tilde{X}_K-\tilde{X}_{i,K}\|>r\}\\
&\leq \mathbb{E}\sum_{i=n+1}^{2n} W_{i,K}(\hat{X}_K)\mathbf{1}\left\{\|\tilde{X}_K-\tilde{X}_{i,K}\|>r, \|\hat{\Lambda}-\Lambda\| \leq \frac{r}{4K}\right\} + \mathbb{P}\left(\|\hat{\Lambda}-\Lambda\|>\frac{r}{4K}\right)\\
&\leq \mathbb{E}\sum_{i=n+1}^{2n} W_{i,K}(\hat{X})\mathbf{1}\left\{\|\hat{X}_K-\hat{X}_{i,K}\|>\frac{r}{2}\right\} + \mathbb{P}\left(\|\hat{\Lambda}-\Lambda\|>\frac{r}{4K}\right). \qquad (5.4)
\end{aligned}$$

Now denote by $\hat{X}_{(i),K}(x)$ the i-th NN of $x \in \mathbb{R}^p$ among $\{\hat{X}_{n+1,K},\cdots,\hat{X}_{2n,K}\}$. Then, since

$$\begin{aligned}
\sum_{i=n+1}^{2n} W_{i,K}(\hat{X}_K)\mathbf{1}\left\{\|\hat{X}_K-\hat{X}_{i,K}\|>\frac{r}{2}\right\} &= \frac{1}{k}\sum_{i=1}^{k}\mathbf{1}\left\{\|\hat{X}_{(i),K}(\hat{X}_K)-\hat{X}_K\|>\frac{r}{2}\right\}\\
&\leq \mathbf{1}\left\{\|\hat{X}_{(k),K}(\hat{X}_K)-\hat{X}_K\|>\frac{r}{2}\right\},
\end{aligned}$$

we can deduce from (5.4), Lemma 5.1 and the fact that $\hat{\Lambda}$ converges to Λ in probability that

$$\mathbb{E}\sum_{i=n+1}^{2n} W_{i,K}(\hat{X}_K)\mathbf{1}\{\|\tilde{X}_K-\tilde{X}_{i,K}\|>r\} \longrightarrow 0.$$

Using (5.3), we get that for all $\varepsilon>0$:

$$\limsup_n \mathbb{E}\sum_{i=n+1}^{2n} W_{i,K}(\hat{X}_K)\left(\varphi(\tilde{X}_K)-\varphi(\tilde{X}_{i,K})\right)^2 \leq \varepsilon^2,$$

hence (5.2) holds. □

Lemma 5.3. *Let $\psi : \mathbb{R}^p \to \mathbb{R}_+$ be a borel function which is bounded by 1. Then, there exists a constant $C > 0$ that only depends on p and such that*

$$\mathbb{E} \sum_{i=n+1}^{2n} W_i(\hat{X})\psi(\tilde{X}_i) \leq C\mathbb{E}\psi(\tilde{X}).$$

Proof By Doob's factorisation Lemma, there exists a borel function $\xi : \mathbb{R}^p \to \mathbb{R}_+$ such that for all $i = n+1, \cdots, 2n$: $\mathbb{E}[\psi(\tilde{X}_i)|\hat{X}_i] = \xi(\hat{X}_i)$. Note that such a function does not depends on i, because the law of the pair $(\tilde{X}_i, \hat{X}_i)$ is independent on i. We let $\mathscr{S} = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ and $\mathscr{E} = \{\hat{X}_{n+1}, \cdots, \hat{X}_{2n}\}$. Then,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=n+1}^{2n} W_i(\hat{X})\psi(\tilde{X}_i)\Big|\mathscr{S}\right] &= \mathbb{E}\left[\mathbb{E}\left[\sum_{i=n+1}^{2n} W_i(\hat{X})\psi(\tilde{X}_i)\Big|\mathscr{S},\mathscr{E},\hat{X}\right]\Big|\mathscr{S}\right] \\
&= \mathbb{E}\left[\sum_{i=n+1}^{2n} W_i(\hat{X})\mathbb{E}\left[\psi(\tilde{X}_i)\Big|\hat{X}_i\right]\Big|\mathscr{S}\right] \\
&= \mathbb{E}\left[\sum_{i=n+1}^{2n} W_i(\hat{X})\xi(\hat{X}_i)\Big|\mathscr{S}\right].
\end{aligned}$$

By Stone's Lemma (e.g. Lemma 6.3 in Györfi et al, 2002), there exists a constant $C > 0$ only depending on p, and such that:

$$\mathbb{E}\left[\sum_{i=n+1}^{2n} W_i(\hat{X})\xi(\hat{X}_i)\Big|\mathscr{S}\right] \leq C\mathbb{E}\left[\xi(\hat{X})\Big|\mathscr{S}\right].$$

This leads to:

$$\begin{aligned}
\mathbb{E}\sum_{i=n+1}^{2n} W_i(\hat{X})\psi(\tilde{X}_i) &= \mathbb{E}\mathbb{E}\left[\sum_{i=n+1}^{2n} W_i(\hat{X})\psi(\tilde{X}_i)\Big|\mathscr{S}\right] \\
&\leq C\mathbb{E}\xi(\hat{X}) = C\mathbb{E}\psi(\tilde{X}),
\end{aligned}$$

by definition of ξ, hence the lemma. □

5.2 Proof of Theorem 2.1

In the sequel, $\tilde{r}$ stands for the function defined for all $x \in \mathbb{R}^p$ by:

$$\tilde{r}(x) = \sum_{i=n+1}^{2n} W_i(x)r(\tilde{X}_i).$$

Fix $\varepsilon > 0$. There exists a continuous function $r' : \mathbb{R}^p \to \mathbb{R}$ with a bounded support such that

$$\mathbb{E}\left(r(\tilde{X}) - r'(\tilde{X})\right)^2 \leq \varepsilon.$$

One may also choose r' so that $0 \leq r' \leq 1$. Since $\sum_{i=n+1}^{2n} W_i(\hat{X}) = 1$, we have by Jensen's inequality:

$$\begin{aligned}\mathbb{E}\left(r(\tilde{X}) - \tilde{r}(\hat{X})\right)^2 &= \mathbb{E}\left(\sum_{i=n+1}^{2n} W_i(\hat{X})\left(r(\tilde{X}) - r(\tilde{X}_i)\right)\right)^2 \\ &\leq \mathbb{E}\sum_{i=n+1}^{2n} W_i(\hat{X})\left(r(\tilde{X}) - r(\tilde{X}_i)\right)^2.\end{aligned}$$

Introducing the continuous function r', we obtain:

$$\begin{aligned}\mathbb{E}\left(r(\tilde{X}) - \tilde{r}(\hat{X})\right)^2 \leq\ & 3\mathbb{E}\left(r(\tilde{X}) - r'(\tilde{X})\right)^2 + 3\mathbb{E}\sum_{i=n+1}^{2n} W_i(\hat{X})\left(r'(\tilde{X}) - r'(\tilde{X}_i)\right)^2 \\ & + 3\mathbb{E}\sum_{i=n+1}^{2n} W_i(\hat{X})\left(r'(\tilde{X}_i) - r(\tilde{X}_i)\right)^2.\end{aligned}$$

According to Lemma 5.3 and by definition of r', we then get:

$$\mathbb{E}\left(r(\tilde{X}) - \tilde{r}(\hat{X})\right)^2 \leq 3\varepsilon(1+C) + 3\mathbb{E}\sum_{i=n+1}^{2n} W_i(\hat{X})\left(r'(\tilde{X}) - r'(\tilde{X}_i)\right)^2,$$

for some constant $C > 0$. Therefore, by Lemma 5.2, we have for all $\varepsilon > 0$:

$$\limsup \mathbb{E}\left(r(\tilde{X}) - \tilde{r}(\hat{X})\right)^2 \leq 3\varepsilon(1+C),$$

and hence

$$\mathbb{E}\left(r(\tilde{X}) - \tilde{r}(\hat{X})\right)^2 \longrightarrow 0. \tag{5.5}$$

The task is now to prove the following property:

$$\mathbb{E}\left(\tilde{r}(\hat{X}) - \hat{r}(\hat{X})\right)^2 \longrightarrow 0.$$

First observe that

$$\mathbb{E}\left(\tilde{r}(\hat{X}) - \hat{r}(\hat{X})\right)^2 = \mathbb{E}\left(\sum_{i=n+1}^{2n} W_i(\hat{X})\left(r(\tilde{X}_i) - Y_i\right)\right)^2.$$

But, if $i, j = n+1, \cdots, 2n$ are different,

$$\begin{aligned}
&\mathbb{E}\left[W_i(\hat{X})(r(\tilde{X}_i) - Y_i)W_j(\hat{X})(r(\tilde{X}_j) - Y_j)\Big| X, X_1, \cdots, X_{2n}, Y_1, \cdots, Y_n\right] \\
&= W_i(\hat{X})W_j(\hat{X})\left(r(\tilde{X}_i) - \mathbb{E}[Y_i|X_i]\right)\left(r(\tilde{X}_j) - \mathbb{E}[Y_j|X_j]\right) \\
&= 0,
\end{aligned}$$

since, by the basic assumption, $\mathbb{E}[Y_i|X_i] = \mathbb{E}[Y_i|\tilde{X}_i] = r(\tilde{X}_i)$. Consequently,

$$\mathbb{E}W_i(\hat{X})(r(\tilde{X}_i) - Y_i)W_j(\hat{X})(r(\tilde{X}_j) - Y_j) = 0,$$

which implies that

$$\mathbb{E}\left(\tilde{r}(\hat{X}) - \hat{r}(\hat{X})\right)^2 = \mathbb{E}\sum_{i=n+1}^{2n} W_i(\hat{X})^2\left(r(\tilde{X}_i) - Y_i\right)^2 \leq \frac{1}{k},$$

because $\sum_{i=n+1}^{2n} W_i(\hat{X}) = 1$, $W_i(\hat{X}) \leq 1/k$ and $|Y| \leq 1$ by assumption. The theorem is now a straightforward consequence of (5.5). □

6 Proof of Theorem 2.2

Recall that we assume here that $k/n \to 0$. We shall make use of the notations of Section 5.1: $\hat{X} = \hat{\Lambda}X$, $\tilde{X} = \Lambda X$ and, for all $i = n+1, \cdots, 2n$: $\tilde{X}_i = \Lambda X_i$. For simplicity, we assume throughout the proof that $\|X\| \leq 1$ and $|Y| \leq 1$. Finally, we denote by $\mathscr{S}$ the sub-sample $\mathscr{S} = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$.

The above proof will borrow and adapt some elements from the proof of Theorem 6.2 in Györfi et al (2002). We first need a lemma.

Lemma 6.1. *If $d \geq 3$, then there exists a constant $C > 0$ such that:*

$$\mathbb{E}\left[\|\hat{X}_{(1)}(\hat{X}) - \hat{X}\|^2 | \mathscr{S}\right] \leq \frac{C}{n^{2/d}},$$

on the event where $\hat{d} \leq d$ and $\|\hat{\Lambda}\| \leq 2\|\Lambda\|$.

Proof We assume throughout the proof that the sub-sample $\mathscr{S}$ is fixed, with $\hat{d} \leq d$ and $\|\hat{\Lambda}\| \leq 2\|\Lambda\|$, and we denote by $\hat{\mu}$ the law of $\hat{X}$ (given $\mathscr{S}$). Since $\hat{d} \leq d$, the support of $\hat{\mu}$ is contained in some vector space of dimension d. For simplicity, we

shall consider that $\hat{\mu}$ is a probability measure on $\mathbb{R}^d$.

We first fix $\varepsilon > 0$. Then,

$$\begin{aligned}
\mathbb{P}(\|\hat{X}_{(1)}(\hat{X}) - \hat{X}\| > \varepsilon | \mathscr{S}) &= \mathbb{E}\left[\mathbb{P}(\|\hat{X}_{(1)}(\hat{X}) - \hat{X}\| > \varepsilon | \mathscr{S}, X) | \mathscr{S}\right] \\
&= \mathbb{E}\left[\mathbb{P}\left(\|\hat{X}_{n+1} - \hat{X}\| > \varepsilon | \mathscr{S}, X\right)^n | \mathscr{S}\right] \\
&= \mathbb{E}\left[\left(1 - \hat{\mu}(B(\hat{X}, \varepsilon))\right)^n | \mathscr{S}\right] \\
&= \int_{\mathbb{R}^d} (1 - \hat{\mu}(B(x, \varepsilon)))^n \, \hat{\mu}(\mathrm{d}x),
\end{aligned}$$

where $B(x,r)$ stands for the Euclidean closed ball in $\mathbb{R}^d$, with center at x and radius r. Since $\|X\| \leq 1$, the support $\mathrm{supp}(\hat{\mu})$ of $\hat{\mu}$ is contained in the ball $B(0, \|\hat{\Lambda}\|)$. Thus, one can find $N(\varepsilon)$ Euclidean balls in $\mathbb{R}^d$ with radius ε, say $B_1, \cdots, B_{N(\varepsilon)}$, such that

$$\mathrm{supp}(\hat{\mu}) \subset \bigcup_{j=1}^{N(\varepsilon)} B_j \text{ and } N(\varepsilon) \leq 2 \frac{\|\hat{\Lambda}\|}{\varepsilon^d}. \tag{6.1}$$

Observe that if $x \in B_j$, then $B_j \subset B(x, \varepsilon)$. Consequently,

$$\begin{aligned}
\mathbb{P}(\|\hat{X}_{(1)}(\hat{X}) - \hat{X}\| > \varepsilon | \mathscr{S}) &\leq \sum_{j=1}^{N(\varepsilon)} \int_{B_j} (1 - \hat{\mu}(B(x, \varepsilon)))^n \, \hat{\mu}(\mathrm{d}x) \\
&\leq \sum_{j=1}^{N(\varepsilon)} \int_{B_j} \left(1 - \hat{\mu}(B_j)\right)^n \hat{\mu}(\mathrm{d}x) \\
&\leq \sum_{j=1}^{N(\varepsilon)} \hat{\mu}(B_j) \left(1 - \hat{\mu}(B_j)\right)^n \\
&\leq \frac{N(\varepsilon)}{n}, \qquad (6.2)
\end{aligned}$$

since $t(1-t)^n \leq 1/n$ when $t \in [0,1]$.

Recall now that $\|X\| \leq 1$ and hence $\|\hat{X}\| \leq \|\hat{\Lambda}\|$. Therefore:

$$\begin{aligned}
\mathbb{E}\left[\|\hat{X}_{(1)}(\hat{X}) - \hat{X}\|^2 | \mathscr{S}\right] &= \int_0^\infty \mathbb{P}(\|\hat{X}_{(1)}(\hat{X}) - \hat{X}\|^2 > \varepsilon | \mathscr{S}) \, \mathrm{d}\varepsilon \\
&= \int_0^{\|\hat{\Lambda}\|^2} \mathbb{P}(\|\hat{X}_{(1)}(\hat{X}) - \hat{X}\| > \sqrt{\varepsilon} | \mathscr{S}) \, \mathrm{d}\varepsilon.
\end{aligned}$$

Using (6.2) and (6.1) lead to the following bound:

$$\begin{aligned}\mathbb{E}\left[\|\hat{X}_{(1)}(\hat{X})-\hat{X}\|^2|\mathscr{S}\right] &\leq \int_0^{\|\hat{\Lambda}\|^2}\min\left(1,\frac{N(\sqrt{\varepsilon})}{n}\right)\mathrm{d}\varepsilon \\ &\leq \int_0^{\|\hat{\Lambda}\|^2}\min\left(1,\frac{2\|\hat{\Lambda}\|}{n\varepsilon^{d/2}}\right)\mathrm{d}\varepsilon.\end{aligned}$$

Since $\|\hat{\Lambda}\| \leq 2\|\Lambda\|$, it is now an easy task to prove that, provided $d \geq 3$,

$$\mathbb{E}\left[\|\hat{X}_{(1)}(\hat{X})-\hat{X}\|^2|\mathscr{S}\right] \leq \frac{C}{n^{2/d}},$$

for some constant $C > 0$, hence the lemma. □

We are now in position to prove Theorem 2.2.

Proof of Theorem 2.2 We shall use the bias-variance decomposition of the following form:

$$\mathbb{E}\left[\left(\hat{r}(\hat{X})-r(\hat{X})\right)^2|\mathscr{S},X\right] = I_1 + I_2, \tag{6.3}$$

where we put, with the notation $\mathscr{S}^W = \mathscr{S} \cup \{X_{n+1},\cdots,X_{2n}\}$:

$$\begin{aligned}I_1 &= \mathbb{E}\left[\left(\hat{r}(\hat{X})-\mathbb{E}\left[\hat{r}(\hat{X})|\mathscr{S}^W,X\right]\right)^2\Big|\mathscr{S},X\right] \\ \text{and } I_2 &= \mathbb{E}\left[\left(\mathbb{E}\left[\hat{r}(\hat{X})|\mathscr{S}^W,X\right]-r(\hat{X})\right)^2\Big|\mathscr{S},X\right].\end{aligned}$$

We first proceed to bound I_1. Let us remark that since, by assumption, $r(\tilde{X}_i) = \mathbb{E}[Y_i|\Lambda X_i] = \mathbb{E}[Y_i|X_i]$, we have:

$$\begin{aligned}\mathbb{E}\left[\hat{r}(\hat{X})|\mathscr{S}^W,X\right] &= \mathbb{E}\left[\sum_{i=n+1}^{2n} W_i(\hat{X})Y_i\Big|\mathscr{S}^W,X\right] = \sum_{i=n+1}^{2n} W_i(\hat{X})\mathbb{E}[Y_i|X_i] \\ &= \sum_{i=n+1}^{2n} W_i(\hat{X})r(\tilde{X}_i).\end{aligned} \tag{6.4}$$

Consequently,

$$\begin{aligned}I_1 &= \mathbb{E}\left[\left(\sum_{i=n+1}^{2n} W_i(\hat{X})\left(Y_i-r(\tilde{X}_i)\right)\right)^2\Big|\mathscr{S},X\right] \\ &= \mathbb{E}\left[\sum_{i=n+1}^{2n} W_i(\hat{X})^2\left(Y_i-r(\tilde{X}_i)\right)^2\Big|\mathscr{S},X\right],\end{aligned}$$

since, as seen in a similar context in the proof of Theorem 2.1,

$$\mathbb{E}\left[W_i(\hat{X})(Y_i - r(\tilde{X}_i))W_j(\hat{X})(Y_j - r(\tilde{X}_j))|\mathscr{S}, X\right] = 0,$$

provided $i, j = n+1, \cdots, 2n$ are different. Using the properties $\sum_{i=n+1}^{2n} W_i(\hat{X}) = 1$, $W_i(\hat{X}) \leq 1/k$ and $|Y| \leq 1$, we conclude that:

$$I_1 \leq \frac{1}{k} \tag{6.5}$$

We now proceed to bound I_2. Since r is a Lipschitz function, there exists a constant $L > 0$ such that $|r(x_1) - r(x_2)| \leq L\|x_1 - x_2\|$ for all $x_1, x_2 \in \mathbb{R}^p$. Then, according to (6.4):

$$\begin{aligned}
I_2 &\leq 2\mathbb{E}\left[\left(\sum_{i=n+1}^{2n} W_i(\hat{X})\left(r(\tilde{X}_i) - r(\hat{X}_i)\right)\right)^2 \Big| \mathscr{S}, X\right] \\
&\quad + 2\mathbb{E}\left[\left(\sum_{i=n+1}^{2n} W_i(\hat{X})\left(r(\hat{X}_i) - r(\hat{X})\right)\right)^2 \Big| \mathscr{S}, X\right] \\
&\leq 2L^2\|\hat{\Lambda} - \Lambda\|^2 + 2\mathbb{E}\left[\left(\frac{1}{k}\sum_{i=1}^{k}\left(r(\hat{X}_{(i)}(\hat{X})) - r(\hat{X})\right)\right)^2 \Big| \mathscr{S}, X\right] \\
&\leq 2L^2\|\hat{\Lambda} - \Lambda\|^2| + 2L^2\mathbb{E}\left[\left(\frac{1}{k}\sum_{i=1}^{k}\|\hat{X}_{(i)}(\hat{X}) - \hat{X}\|\right)^2 \Big| \mathscr{S}, X\right], \qquad (6.6)
\end{aligned}$$

where we used the facts that $\|X\| \leq 1$ and $\sum_{i=n+1}^{2n} W_i(\hat{X}) = 1$. We now let $\tilde{n} = [n/k]$, and we split the sub-sample $\{\tilde{X}_1, \cdots, \tilde{X}_{k\tilde{n}}\}$ into k sub-samples $Z_1, \cdots, Z_k$ of size $\tilde{n}$, with:

$$Z_i = \{\tilde{X}_{i\tilde{n}+1}, \cdots \tilde{X}_{(i+1)\tilde{n}}\},\ i = 1, \cdots, k.$$

For each sample Z_i, we denote by $Z_i^{(1)}$ the closest element of Z_i from $\hat{X}$ (ties being considered as usual). Then,

$$\sum_{i=1}^{k}\|\hat{X}_{(i)}(\hat{X}) - \hat{X}\| \leq \sum_{i=1}^{k}\|Z_i^{(1)} - \hat{X}\|.$$

Jensen's Inequality and (6.6) then give

$$\mathbb{E}\left[I_2|\mathscr{S}\right] \leq 2L^2\|\hat{\Lambda} - \Lambda\|^2 + \frac{2L^2}{k}\sum_{i=1}^{k}\mathbb{E}\left[\|Z_i^{(1)} - \hat{X}\|^2\Big|\mathscr{S}\right].$$

Therefore, on the event where $\hat{d} \leq d$ and $\|\hat{\Lambda}\| \leq 2\|\Lambda\|$, we have by Lemma 6.1:

$$\mathbb{E}[I_2|\mathscr{S}] \leq 2L^2\|\hat{\Lambda} - \Lambda\|^2 + \frac{2L^2C}{\tilde{n}^{2/d}},$$

for some constant $C > 0$. Since $k/n \to 0$, there exists a constant $\kappa > 0$ such that $\tilde{n} \geq \kappa n/k$. Hence, on the event where $\hat{d} \leq d$ and $\|\hat{\Lambda}\| \leq 2\|\Lambda\|$,

$$\mathbb{E}[I_2|\mathscr{S}] \leq 2L^2\|\hat{\Lambda} - \Lambda\|^2 + 2\kappa^{-2/d}L^2C\left(\frac{k}{n}\right)^{2/d}.$$

By (6.5) and (6.3), we then deduce that for some constant $C' > 0$:

$$\mathbb{E}\left[\left(\hat{r}(\hat{X}) - r(\hat{X})\right)^2 \middle| \mathscr{S}\right] \leq \frac{1}{k} + C'\|\hat{\Lambda} - \Lambda\|^2 + C'\left(\frac{k}{n}\right)^{2/d},$$

on the event where $\hat{d} \leq d$ and $\|\hat{\Lambda}\| \leq 2\|\Lambda\|$. Noticing that $\|\hat{\Lambda} - \Lambda\| > \|\Lambda\|$ when $\|\hat{\Lambda}\| > 2\|\Lambda\|$, and since $|\hat{r}(\hat{X})| \leq 1$, $|r(\hat{X})| \leq 1$, we obtain:

$$\begin{aligned}
\mathbb{E}\left(\hat{r}(\hat{X}) - r(\hat{X})\right)^2 &= \mathbb{E}\mathbb{E}\left[\left(\hat{r}(\hat{X}) - r(\hat{X})\right)^2 \middle| \mathscr{S}\right] \\
&\leq \frac{1}{k} + C'\mathbb{E}\|\hat{\Lambda} - \Lambda\|^2 + C'\left(\frac{k}{n}\right)^{2/d} + \mathbb{P}(\|\hat{\Lambda} - \Lambda\| > \|\Lambda\|) \\
&\quad + \mathbb{P}(\hat{d} > d) \\
&\leq \frac{1}{k} + \left(C' + \frac{1}{\|\Lambda\|^2}\right)\mathbb{E}\|\hat{\Lambda} - \Lambda\|^2 + C'\left(\frac{k}{n}\right)^{2/d} + \mathbb{P}(\hat{d} > d),
\end{aligned}$$

using the Markov Inequality. Finally, by the Lipschitz property of r,

$$\mathbb{E}\left(r(\hat{X}) - r(\tilde{X})\right)^2 \leq L^2\mathbb{E}\|\hat{\Lambda} - \Lambda\|^2.$$

The last 2 inequalities give the result since, by the basic assumption, $r(\tilde{X}) = \mathbb{E}[Y|X]$. □

7 Proof of Corollary 3.1

The proof of Corollary 3.1 is straightforward from Corollary 2.2 and Lemmas 7.1 and 7.2 below.

Lemma 7.1. *We have:*

$$\mathbb{P}(\hat{d} > d) \leq \frac{p}{\tau^2}\,\mathbb{E}\|\hat{M} - M\|^2.$$

Proof Let $\mathscr{N} = \{j = 1, \cdots, p : \lambda_j \neq 0\}$, and recall that $\mathrm{card}(\mathscr{N}) = d$. If $\hat{d} > d$, we then have

$$1 \leq \sum_{j \in \mathscr{N}} \big(\mathbf{1}\{|\hat{\lambda}_j| \geq \tau\} - 1\big) + \sum_{j \notin \mathscr{N}} \mathbf{1}\{|\hat{\lambda}_j| \geq \tau\} \leq \sum_{j \notin \mathscr{N}} \mathbf{1}\{|\hat{\lambda}_j| \geq \tau\}.$$

Thus, we deduce the inequality:

$$\mathbb{P}(\hat{d} > d) \leq p \max_{j \notin \mathscr{N}} \mathbb{P}(|\hat{\lambda}_j| \geq \tau). \tag{7.1}$$

Fix $j \notin \mathscr{N}$. Since $\hat{M}$ and M are symmetric, the indexation of the eigenvalues implies that $\|\hat{M} - M\| \geq |\hat{\lambda}_j - \lambda_j| = |\hat{\lambda}_j|$. Consequently, when $|\hat{\lambda}_j| \geq \tau$, we have $\|\hat{M} - M\| \geq \tau$. Therefore:

$$\mathbb{P}(|\hat{\lambda}_j| \geq \tau) \leq \mathbb{P}(\|\hat{M} - M\| \geq \tau) \leq \frac{1}{\tau^2}\,\mathbb{E}\|\hat{M} - M\|^2.$$

The lemma is now a straightforward consequence of (7.1). □

Our next task is to bound the quantity $\mathbb{E}\|\Lambda - \hat{\Lambda}\|^2$. For this purpose, we recall the following classical fact (e.g. see Kato, 1966): for any symmetric matrix $A \in \mathscr{M}_p(\mathbb{R})$, let $v_i(A)$ be the normalized eigenvector associated with the i-th largest eigenvalue. If it is a simple eigenvalue, then there exists $\delta_A > 0$ such that for any symmetric matrix $A' \in \mathscr{M}_p(\mathbb{R})$ with $\|A - A'\| \leq \delta_A$:

$$\|v_i(A) - v_i(A')\| \leq C_0 \|A - A'\|, \tag{7.2}$$

for some constant $C_0 > 0$ that only depends on A.

Lemma 7.2. *Assume that the non-null eigenvalues of M have multiplicity 1. Then, there exists a constant $C > 0$ such that:*

$$\mathbb{E}\|\hat{\Lambda} - \Lambda\|^2 \leq \frac{C}{\tau^2}\,\mathbb{E}\|\hat{M} - M\|^2.$$

Proof We let

$$\mathscr{N} = \{j = 1, \cdots, p : \lambda_j \neq 0\} \quad \text{and} \quad \hat{\mathscr{N}} = \{j = 1, \cdots, p : |\hat{\lambda}_j| \geq \tau\}.$$

Writting $\hat{M} = M + (\hat{M} - M)$, we deduce from (7.2) that, provided $\|\hat{M} - M\| \leq \delta_M$:

$$\max_{j \in \mathscr{N}} \|\hat{v}_j - v_j\| \leq C \|\hat{M} - M\|, \tag{7.3}$$

Here, and in the following, C is a positive constant whose value may change from line to line. Since $\|v_j\| = \|\hat{v}_j\| = 1$ for all $j = 1, \cdots, p$, we have:

$$\begin{aligned}
&\mathbb{E}\|\hat{\Lambda} - \Lambda\|^2 \mathbf{1}\{\|\hat{M} - M\| \leq \delta_M\} \\
&\quad = \mathbb{E} \sum_{j \in \mathscr{N} \cup \hat{\mathscr{N}}} \|\hat{v}_j - v_j\|^2 \mathbf{1}\{\|\hat{M} - M\| \leq \delta_M\} \\
&\quad \leq \sum_{j \in \mathscr{N}} \mathbb{E}\|\hat{v}_j - v_j\|^2 \mathbf{1}\{\|\hat{M} - M\| \leq \delta_M\} + \sum_{j=1}^{p} \mathbb{E}\|\hat{v}_j - v_j\|^2 \mathbf{1}\{\mathscr{N} \cup \hat{\mathscr{N}} \neq \mathscr{N}\} \\
&\quad \leq C\,\mathbb{E}\|\hat{M} - M\|^2 + C\,\mathbb{P}(\mathscr{N} \cup \hat{\mathscr{N}} \neq \mathscr{N}),
\end{aligned}$$

according to (7.3). Moreover, since $\|\hat{M} - M\| \geq |\hat{\lambda}_j - \lambda_j|$ for all j because M and $\hat{M}$ are symmetric matrices, we have:

$$\begin{aligned}
\mathbb{P}(\mathscr{N} \cup \hat{\mathscr{N}} \neq \mathscr{N}) &= \mathbb{P}\left(\exists j = 1, \cdots, p : |\hat{\lambda}_j| \geq \tau \quad \text{and} \quad \lambda_j = 0\right) \\
&\leq \mathbb{P}(\|\hat{M} - M\| \geq \tau) \leq \frac{1}{\tau^2}\,\mathbb{E}\|\hat{M} - M\|^2.
\end{aligned}$$

Combining the next two inequalities gives:

$$\mathbb{E}\|\hat{\Lambda} - \Lambda\|^2 \mathbf{1}\{\|\hat{M} - M\| \leq \delta_M\} \leq C\,\mathbb{E}\|\hat{M} - M\|^2 + \frac{C}{\tau^2}\,\mathbb{E}\|\hat{M} - M\|^2.$$

Since $\|\hat{\Lambda} - \Lambda\| \leq 2p$, we also have:

$$\begin{aligned}
\mathbb{E}\|\hat{\Lambda} - \Lambda\|^2 \mathbf{1}\{\|\hat{M} - M\| > \delta_M\} &\leq C\,\mathbb{P}(\|\hat{M} - M\| > \delta_M) \\
&\leq C\,\mathbb{E}\|\hat{M} - M\|^2.
\end{aligned}$$

Therefore, since $\tau \leq 1$:

$$\mathbb{E}\|\hat{\Lambda} - \Lambda\|^2 \leq \frac{C}{\tau^2}\,\mathbb{E}\|\hat{M} - M\|^2,$$

for some constant $C > 0$, that only depends on M and p. □

REFERENCES

Cook, R.D. (1994). On the Interpretation of Regression Plot. *Journal of the American Statistical Association* **89**, 177-189.

Cook, R.D. (1996). Graphics for Regression with a Binary Response. *Journal of the American Statistical Association* **91**, 983-992.

Cook, R.D. (1998). *Regression Graphics*. Wiley, New-York NY.

Cook, R.D. and Li, B. (2002). Dimension Reduction for Conditional Mean in Regression, *The Annals of Statistics* **30**, 455-474.

Cook, R.D. and Ni, L. (2005). Sufficient Dimension Reduction via Inverse Regression: A Minimum Discrepancy Approach. *Journal of the American Statistical Association* **100**, 410-428.

Cook, R.D. and Weisberg, S. (1991). Discussion of "Sliced Inverse Regression for Dimension Reduction". *Journal of the American Statistical Association* **86**, 316-342.

Cook, R.D. and Weisberg, S. (1999). Graphs in Statistical Analysis: Is the Medium the Message? *The American Statistician* **53**, 29-37.

Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New-York NY.

Hall, P. and Li, K.C. (1993). On Almost Linearity of Low-Dimensional Projections from High-Dimensional Data. *The Annals of Statistics* **21**, 867-889.

Härdle, W. and Stoker, T.M. (1989). Investigating Smooth Multiple Regression by the Method of Average Derivative. *Journal of the American Statistical Association* **84**, 986-995.

Ibragimov, I.A. and Khasminskii, R.Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New-York NY.

Kato, T. (1966). *Perturbation Theory for Linear Operators*. Springer-Verlag, New-York NY.

Li, K.C. (1991). Sliced Inverse Regression for Dimension Reduction (with Discussion). *Journal of the American Statistical Association* **86**, 316-342.

Li, K.C. (1992). On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma. *Journal of the American Statistical Association* **87**, 1025-1039.

Saracco, J. (2005). Asymptotics for Pooled Marginal Slicing Estimator Based on SIRα. *Journal of Multivariate Analysis* **96**, 117-135.

Xia, Y., Tong, H., Li, W.K. and Zhu, L.-X. (2002). An Adaptative Estimation of Dimension Reduction Space. *Journal of the Royal Statistical Society, Ser. B* **64**, 1-28.

Ye, Z. and Weiss, R.E. (2003). Using the Bootstrap to Select One of a New Class of Dimension Reduction Methods. *Journal of the American Statistical Association* **98**, 968-979.

Zhu, L., Miao, B. and Peng, H. (2006). On Sliced Inverse Regression with High-Dimensional Covariates. *Journal of the American Statistical Association* **101**, 630-643.

Zhu, Y. and Zeng P. (2006). Fourier Methods for Estimating the Central Subspace and the Central Mean Subspace in Regression. *Journal of the American Statistical Association* **101**, 1638-1651.